(* TWO-WAY *)

18x

Going Further With Our Fit

(still East Coast)

		Exh	Page
Introduction	18x-1		
			A
		1	18x-2/3
18xA. Individualized lines	18x-1	2	18x-5
- general remarks	18x-12	3	18x-6
		4	18x-7
18xB. A further step	18x-12	5	18x-8
		6	18x-10
18xC. How far should we go?	18x-19	7	18x-11
degrees of freedom,	18x-21	8	18x-13
- looking deeper	18x-25		
			B
		9	18x-14/15
**********		10	18x-17
		11	18x-18
18xH. How far have we come?	18x-28	12	18x-20
			C
**********		13	18x-22
		14	18x-24
		15	18x-27
			H
			P

18x

Going Further With Our Fit
(still East Coast)

As we closed the analysis of chapter 18, we made it clear that there were further steps of fitting that we might like to take.

This chapter takes some of these steps, all of them kinds of steps that could be attempted for any two-way table, since they use row effects and column effects as their essential building blocks.

18xA. Individualized lines

Before we fitted the additional term, the residuals from the East Coast monthly temperatures were strongly structured. Almost every place showed residuals of one sign in summer and of the other sign in winter. Fitting just one more constant, in a way that twisted each place line in the two-way plot, got rid of much of this systematic pattern. What remained, as we noted in discussing exhibit 30 of chapter 18, follows a rather irregular pattern.

If we are to go further and adjust the size of the January-to-July-to-January swing separately for each place, we need to twist each place line separately. An easy way to assess reasonable amounts of twist begins by looking at the three hot months and the three cold months, as in exhibit 1. Making this additional fit lowers the midspread of the residuals from 1.8h to 1.4h. Clearly we have improved the fit.

exhibit 1 of chapter 18x

Calculation of Individual Twists in
the Two-way Plot
(see exhibits 25 and 30 of chapter 18)

A) MEDIANS of THREE RESIDUALS

	Laredo	Baton Rouge	Atlanta	Wash'n	Boston	Port	Caribou	eff
Hot months	0.8	-1.5	-0.1	1.2	0	-0.4	-0.1	18.5
Cold months	-2.9	1.2	0.7	-0.7	0.9	0.9	-1.1	-18.7
Difference	3.7	-2.7	-0.8	1.9	-0.9	-1.3	1.0	37.2
Ratio*	.10	-.075	-.02	.05	-.025	-.035	.03	(1.00)

B) ADJUSTMENTS to FIT

	L	BR	A	W	B	P	C	(eff**)
Jan	-2.0	1.5	.4	-1.0	0.5	0.7	-0.6	
Feb	-1.9	1.4	.4	-1.0	0.5	0.7	-0.6	
Mar	-1.1	0.8	.2	-0.6	0.3	0.4	-0.4	
Apr	-0.1	0.1	0	0	0	0	0	
May	0.8	-0.6	-.2	0.4	-0.2	-0.3	0.2	
June	1.6	-1.2	-.3	0.8	-0.4	-0.6	0.5	
July	2.0	-1.5	-.4	1.0	-0.5	-0.7	0.6	
Aug	1.8	-1.4	-.3	0.9	-0.4	-0.6	0.5	
Sept	1.3	-1.6	-.2	0.6	-0.3	-0.4	0.4	
Oct	0.2	-0.2	0	0.1	0	0.0	0.0	
Nov	-0.9	.7	.2	-0.4	0.2	0.3	-0.2	
Dec	-1.8	1.4	.4	-0.9	0.4	0.6	-0.5	

* Ratio = "Difference for place" as a fraction of "difference for effect".

**From exhibit of chapter 18.

(Cont'd)

Examples: -2.0 = (.10)( ).

exhibit 1 (Cont'd)

C) NEW RESIDUALS

	L	BR	A	W	B	P	C
Jan	-.9	-.3	.3	1.2	.4	.2	-.5
Feb	2.1	1.6	.8	.3	-1.5	-1.1	-1.8
Mar	2.4	.5	-.2	.1	-1.7	-.1	-.9
Apr	2.6	.3	.7	-.6	-2.0	-.9	-.6
May	.8	.4	1.0	0	-1.4	-1.4	.5
June	-.8	.6	1.6	.4	-.5	-1.3	-.6
July	-1.8	-.5	.2	.4	.5	.4	0
Aug	-.8	-.1	.2	-.3	.9	.2	-.6
Sept	-.6	1.0	1.2	.1	.1	-.9	-1.2
Oct	-.1	-.1	-1.4	-.7	1.2	.5	1.2
Nov	-2.2	-.8	-3.4	-.2	2.4	2.4	3.3
Dec	-1.1	-1.2	-1.6	.1	1.1	1.1	1.6

D) NEW RESIDUALS AS A BATCH -- unit = 0.1

|to|		
1	-3	4
	-2	
3	-2	02
8	-1	58786
16	-1	14432412
32	-0	9886586759959666
42	-0	1311203210
(16)	0	3432231441141212
26	0	856875955
17	1	20022211
9	1	666
6	2	1444
2	2	6
1	3	3

42 + 16 + 26 = 84, hinges at 21h

Hinges:	- .8	and	+ .6h
Midspread:			1.4h
Sides:	-2.2h	and	2.1
Outside:	1	and	5
(count/20):			(4)

$$\text{clarity} = \frac{89.5-9.2}{1.45} = 55.4$$

(see exhibit 6 for extremes of fit)

Our next task is to plot the fit so that we can look at it. Exhibit 2 shows, at an enlarged scale, what changes we need to make in the Laredo line in order that the two-way plot will show the new fit. The amounts of additional shift -- here -2.0 in January and +2.0 in July -- come from panel B of exhibit 1.

Now that we see how to handle the twist of individual lines, we can make up exhibit 3, the construction plot for the new two-way plot, exhibit 4. We still have a reasonably eyeable and understandable plot.

Exhibit 5 shows the coded, but uncondensed residuals. We did not condense residuals here because we knew from the numbers of exhibit 1 (panel C) that spring and fall months with similar temperatures were behaving quite differently. This plot shows a considerable amount of structure. Again we can interpret it in two ways:

- for each individual month it seems to be mainly a matter of "North of Washington one way, South of Washington the other".

- for each individual place we see -- especially in panel C of exhibit 1 -- tendencies to bulge up or down at particular seasons of the year.

Before we pursue this instance further, we ought to stop and take stock, both generally and with regard to our East Coast monthly temperatures.

exhibit 2 of chapter 18x

Three Successive Locations of the Laredo Line

(see exhibit 30 of chapter 18)

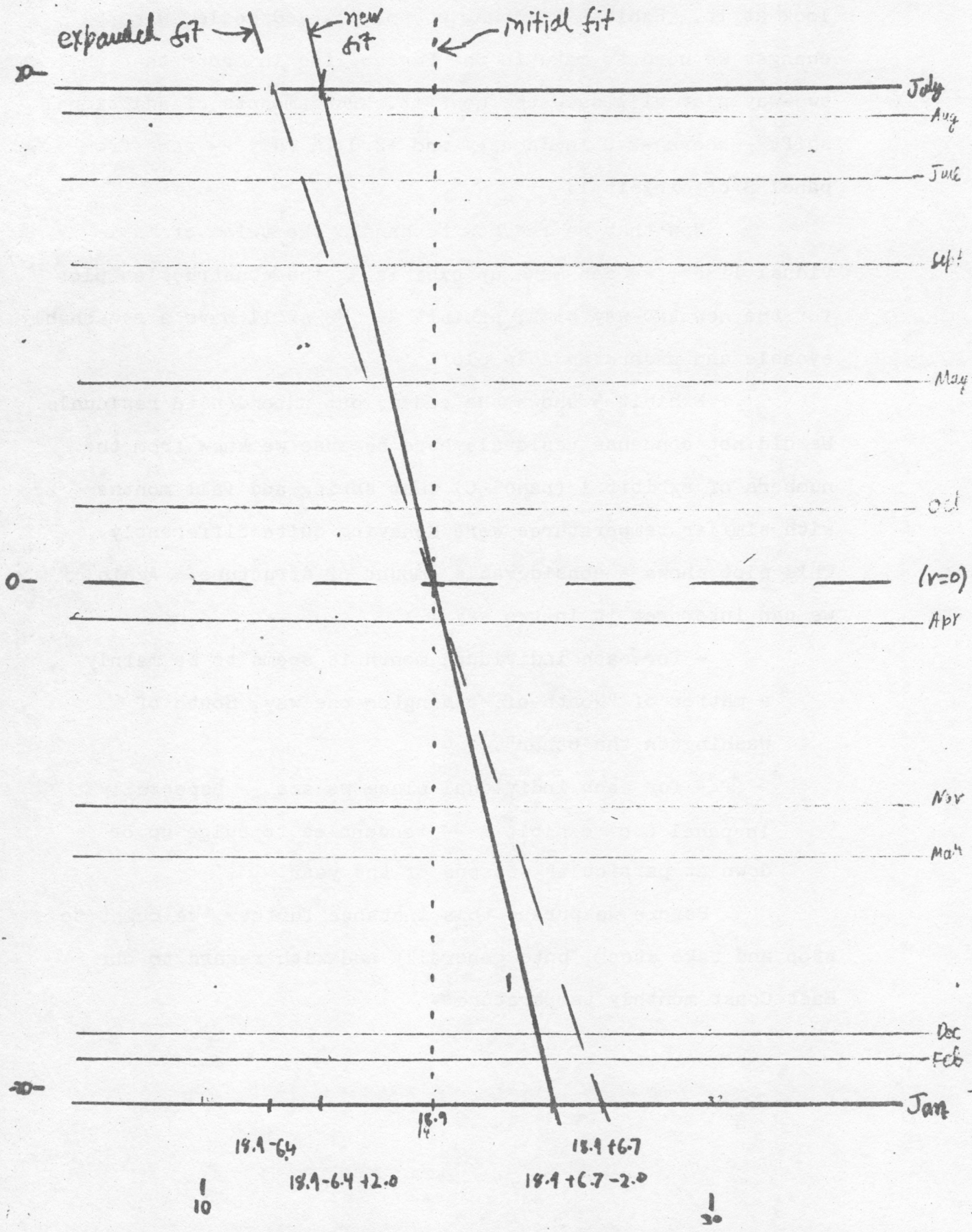


exhibit 3 of chapter 18x

Construction Plot for Two-way Plot
Individually Fitted Place Lines
(see exhibit 1)

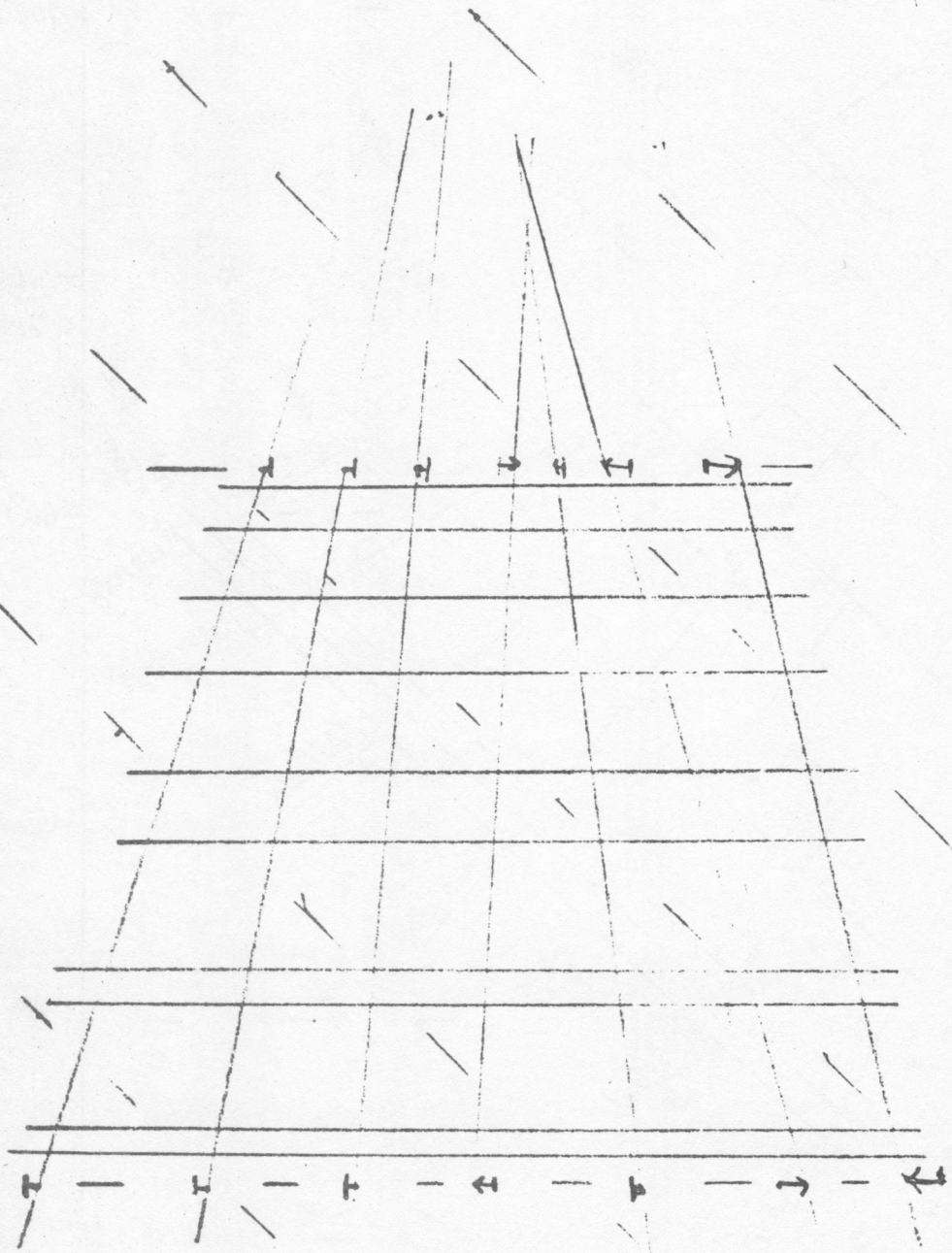


exhibit 4 of chapter 18x

Two-way Plot With Individually Twisted Place Lines

(see exhibit 3)

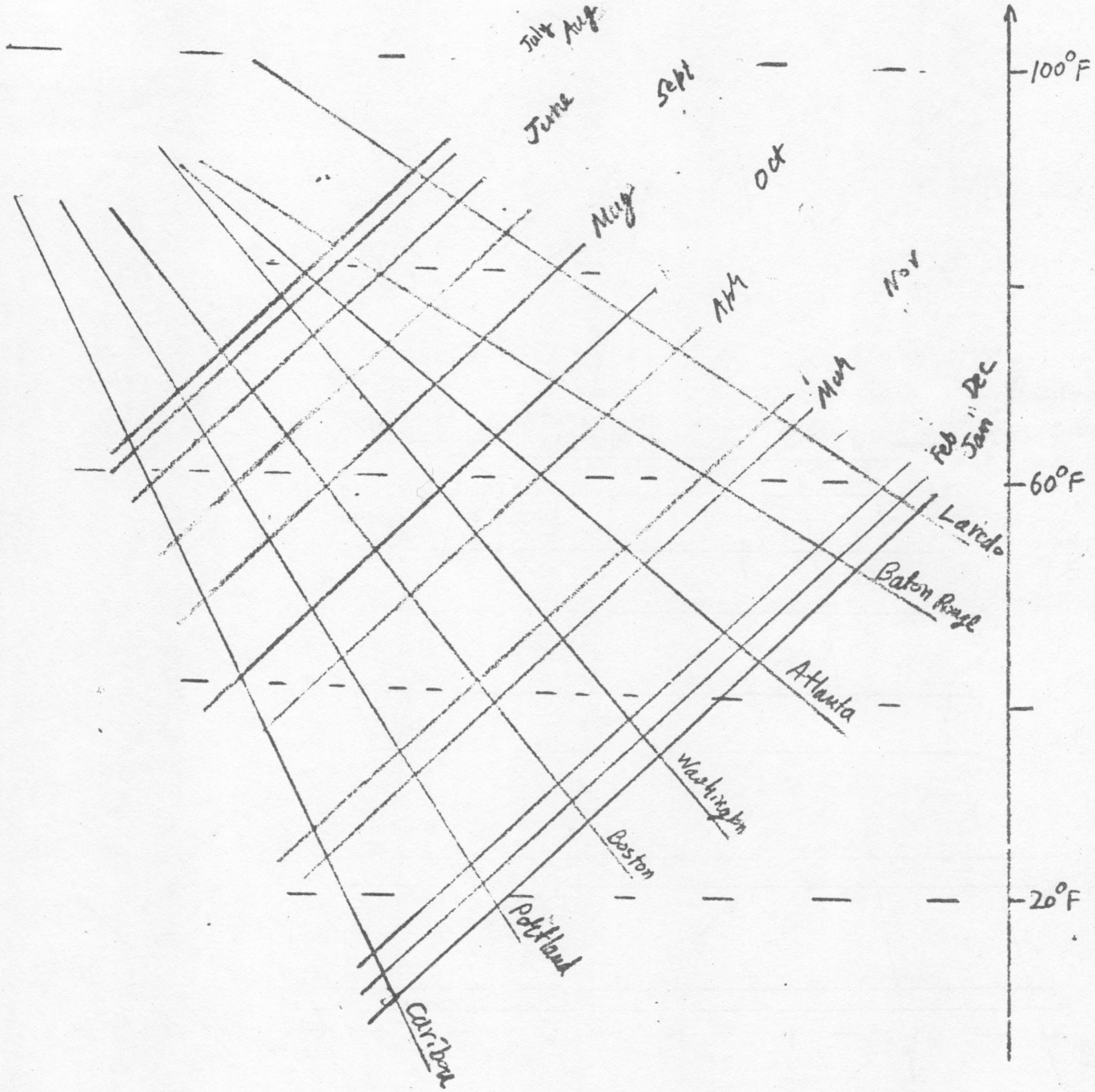
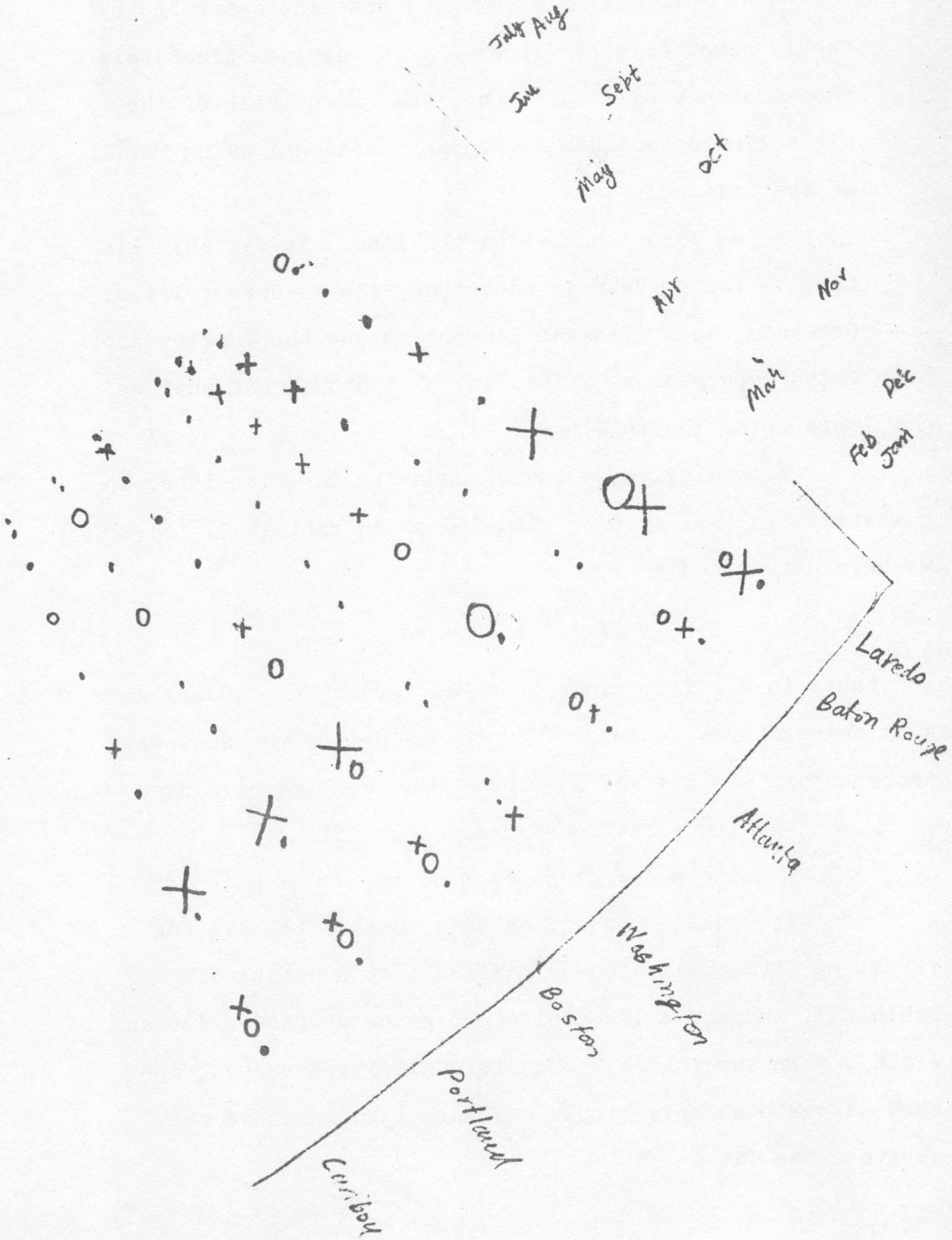


exhibit 5 of chapter 18x

Uncondensed Residuals after Fit
by Separate Place Lines
(see exhibits 1 and 4)



In the example -- as exhibit 4 tells us if we question it carefully -- one way to describe our fit is as follows:

- we have a single seasonal pattern, described by the 12 parallel month lines. We apply a separately chosen amount of it at each place. (The tilt of the place line determines how large a seasonal swing we fit at that place.)

- for each place we have a line. To fix this line takes two numbers, since the line is unrestricted. One easy way to fix the line is to use the January and July temperatures. (This is, of course, just what we were doing in exhibit 2.)

Accordingly, we could summarize our fit either in numbers as in exhibit 6 or visually as in exhibit 7. Clearly we have determined

$$7 + 7 + 10 = 24$$

constants in our fit, since it takes 7 each for January and July temperatures -- and 10 for the values of the seasonal pattern that are not set at 0.000 or 1.000. We have used

- quite a few constants in our fit.

- quite a few steps to find it.

It would, however, be quite unfair to call the resulting fit complicated or difficult to understand. Exhibit 7, in particular, while not doing as good a job as exhibit 4 in showing us the individual fitted values, does show us quite clearly simple elements into which we can separate the fit.

exhibit 6 of chapter 18x

Numerical Summary of the Present Fit

A) EXTREME FITTED TEMPERATURES by PLACE -- see exhibit 22 and panel C of chapter 18 and exhibit 1 above. Calculated from actual values and residuals.

	Laredo	Baton Rouge	Atlanta	Wash'n	Boston	Portl	Caribou
January	58.5	52.7	44.3	35.0	28.7	20.5	9.2
July	89.5	81.6	79.3	76.9	72.7	67.4	64.0

Examples: 58.5 = 57.6 - (-0.9), 89.5 = 87.7 - (-1.8) .

B) SEASONAL PATTERN -- stretch to fit

	(as effects)	(same + 20.5)	handy form
Jan	(-20.5)	(0)	0.000
Feb	(-18.7)	(1.8)	.045
Mar	(-10.7)	(9.8)	.244
Apr	(- 1.4)	(19.1)	.478
May	(7.9)	(28.4)	.710
June	(15.9)	(36.4)	.910
July	(19.6)	(40.1)	1.000
Aug	(18.5)	(39.0)	.975
Sept	(12.6)	(33.1)	.826
Oct	(2.9)	(23.4)	.585
Nov	(-8.1)	(12.4)	.292
Dec	(-17.7)	(2.8)	.070

Examples: -20.5, -18.7, etc. from panel B of exhibit 22 of chapter 18.

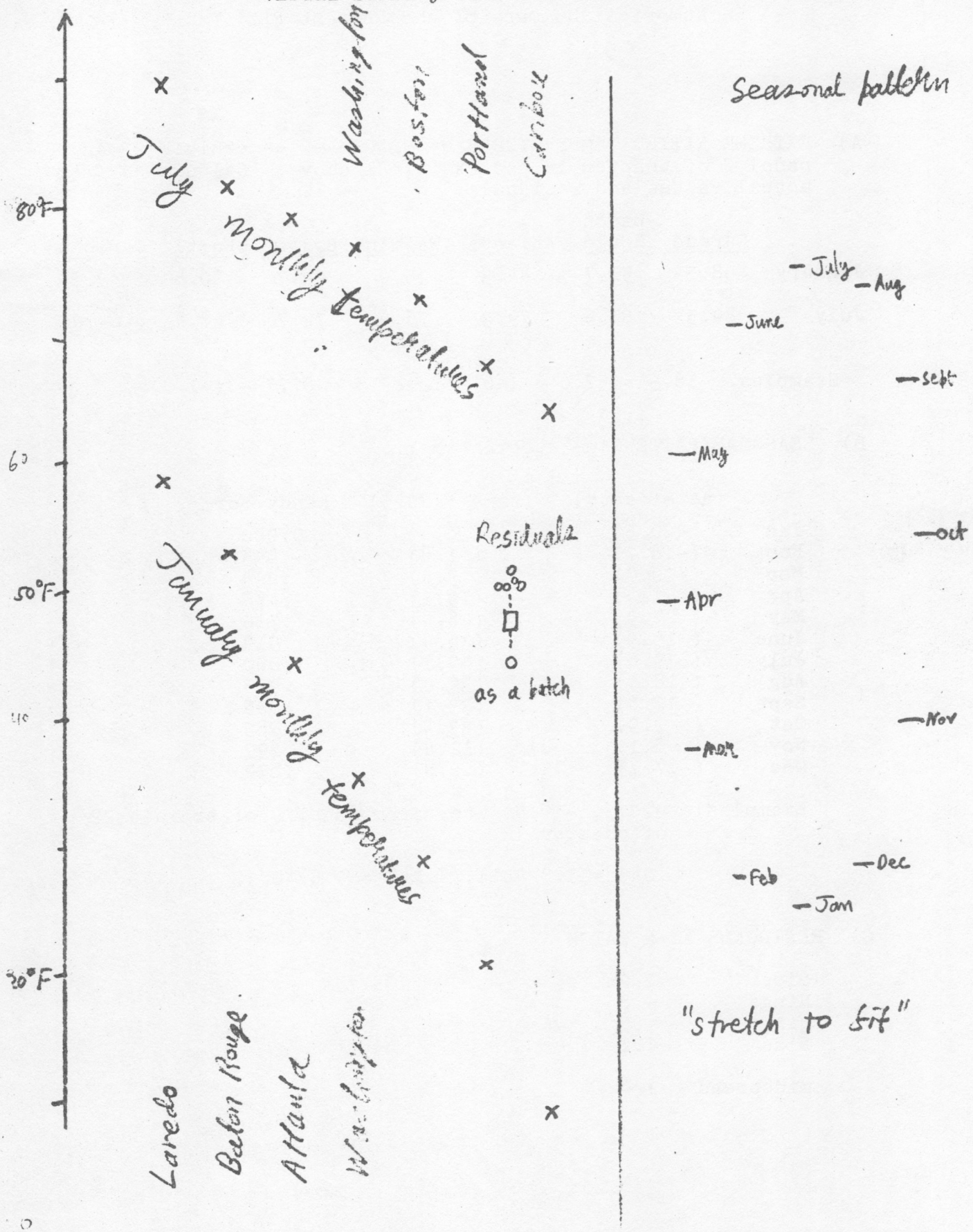
.045 = 1.8/40.1, .244 = 9.8/40.1.

C) RESIDUALS as a BATCH

Side	-2.2h
Hinge	-.8
Hinge	.6h
Side	2.1
midspread	1.4h

exhibit 7 of chapter 18x

Visual Summary of the Present Fit



* general remarks *

While this example was specially easy to understand, we could do to any two-way table any and all of the things we have done to these East Coast monthly temperatures. And, however much or little adjustment seemed to be needed, the resulting two-way plots would have involved straight lines with the same characteristics -- namely those listed in exhibit 8 for easy reference.

Note that we have not been explicit about "(5)", which covers the general case of any row lines and any column lines. This is because we do not know any really simple description of the most general fit that can be described by a two-way plot with two arbitrary sets of lines.

The basic message is simple: we can go a long way with fits that we can see with the aid of two-way plots using straight lines.

18xB. A further step

When last seen, the residuals of the East Coast monthly temperatures still had considerable structure. They still asked for further fitting.

The simplest thing to fit -- either for each month or for those months where it matters -- is a simple distinction between north of Washington and south of Washington, a simple multiple of the values set out in panel A of exhibit 9. The rest of exhibit 9 carries out the fitting of these additional terms.

exhibit 8 of chapter 18x

Characteristics of Lines in Two-way Plots

	What is fitted	Row lines	Column lines
(1)	Row effects and column effects	Parallel	Parallel
(2)	(1) and a multiple of (row eff) × (col eff)		
	either →	Parallel	Radiating
	or →	Radiating	Parallel
(3)	(2) and separate multiples of row effects for each column	Parallel	Any lines
(4)	(2) and separate multiples of column effects for each row	Any lines	Parallel
(5)	Something more general than both (3) and (4).	Any lines	Any lines
(6)	Still more general --	(curves, or broken lines, not all straight lines) --	

exhibit 9 of chapter 18x

Further Fitting to East Coast
Monthly Temperatures
(see exhibit 1)

A) The CARRIER -- multiples of this are to be fit

Place	Value
Laredo	+1
Baton Rouge	+1
Atlanta	+1
Washington	0
Boston	-1
Portland	-1
Caribou	-1

Note: It is convenient to have a name for that whose multiples we are prepared to fit. We shall use the word "carrier" for this purpose.

B) MEANS OF THREE RESIDUALS

	S of Wash'n	N of Wash'n	Diff	$\frac{\text{Diff}}{2}$
Jan	-.3	0	-.3	-.2
Feb	1.5	-1.5	3.0	1.5
Mar	0.9	-0.9	1.8	.9
Apr	1.2	-1.2	2.4	1.2
May	0.7	-0.7	1.4	.7
June	0.5	-0.8	1.3	.6
July	-0.7	0.3	-1.0	-.5
Aug	-0.2	0.2	-0.4	-.2
Sept	0.5	-0.7	1.2	.6
Oct	-0.5	0.9	-1.4	-.7
Nov	-2.1	2.7	-4.8	-2.4
Dec	-1.3	1.3	-2.6	-1.3

(Cont'd)

exhibit 9 (Cont'd)

C) NEW RESIDUALS

	Laredo	Baton Rouge	Atlanta	Wash'n	Boston	Portland	Caribou
Jan	-.7	-.1	.5	1.2	.2	.0	-.7
Feb	.6	.1	-.7	.3	0	.4	-.3
Mar	1.5	-.4	-1.1	.1	-.8	.8	.0
Apr	1.4	-.9	-.5	-.6	-.8	.3	.6
May	.1	-.3	.3	0	-.7	-.7	1.2
June	-1.4	.0	1.0	.4	.1	-.7	.0
July	-1.3	.0	.7	.4	.0	-.1	-.5
Aug	-.6	.1	.4	-.3	.7	.0	-.8
Sept	-1.2	.4	.6	.1	.7	-.3	-.6
Oct	.6	.6	-.7	-.7	.5	-.2	.5
Nov	.2	1.6	-1.0	-.2	0	0	.9
Dec	.2	.1	-.3	.1	-.2	-.2	.4

Examples: -.7=-.2-(-.2), -.1=-.3-(-.2),...-.7=-.5-(+.2),.....
.4=1.7-(+1.3)

D) NEW RESIDUALS AS A BATCH -- unit = 0.1

to		
2	-1	43
5	-1	120
17	-0	777898777877
23	-0	456566
34	-0	13313322322
(11)	±0	00000000000
39	+0	21313131112211
25	+0	56464444666554
11	+0	87779
6	+1	220
3	+1	54
1	+1	6

34 + 11 + 39 = 84, hinges at 21h

Hinges:	- .5	and	+ .4
Midspread:		.9	
Sides:	-1.4	and	+1.3
Outside:	0	and	3
(count/20):		(4)	

$$\text{clarity} = \frac{89.0-9.4}{0.9} = 89$$

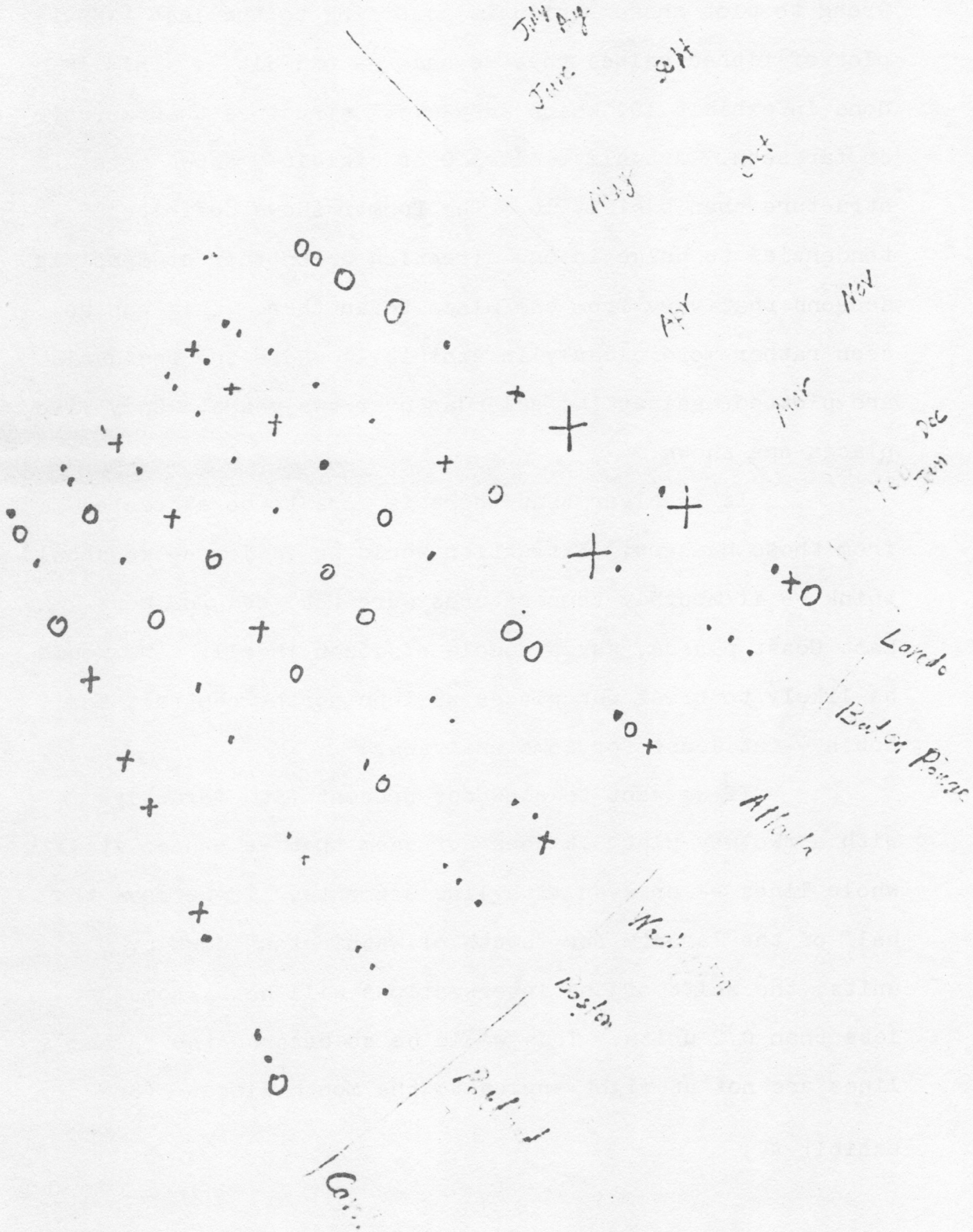
For the moment, we will skip over the problem of two-way plotting the new fit and turn to the even more interesting question of plotting the residuals. We have made only small changes in the fit. We will not go far wrong to plot these residuals according to the last two-way plot of fitted values that we have -- exhibit 4. This is done in exhibit 10, which shows less structure than we have so far seen. Actually, panel C of exhibit 9 shows more structure than exhibit 10. The former shows definite tendencies to bulge in one direction or another at specific seasons that vary from one place to another. This can be seen rather more clearly in exhibit 11, where the residuals are plotted against the calendar over two years. Only five places are shown.

It is clear that there is more to be extracted from these numbers. Extraction would be easier -- we should think -- if monthly temperatures were analyzed for more East Coast places, say a couple of dozen in all. (We would be likely to break our places up into north, central, and south -- at least for some analyses.)

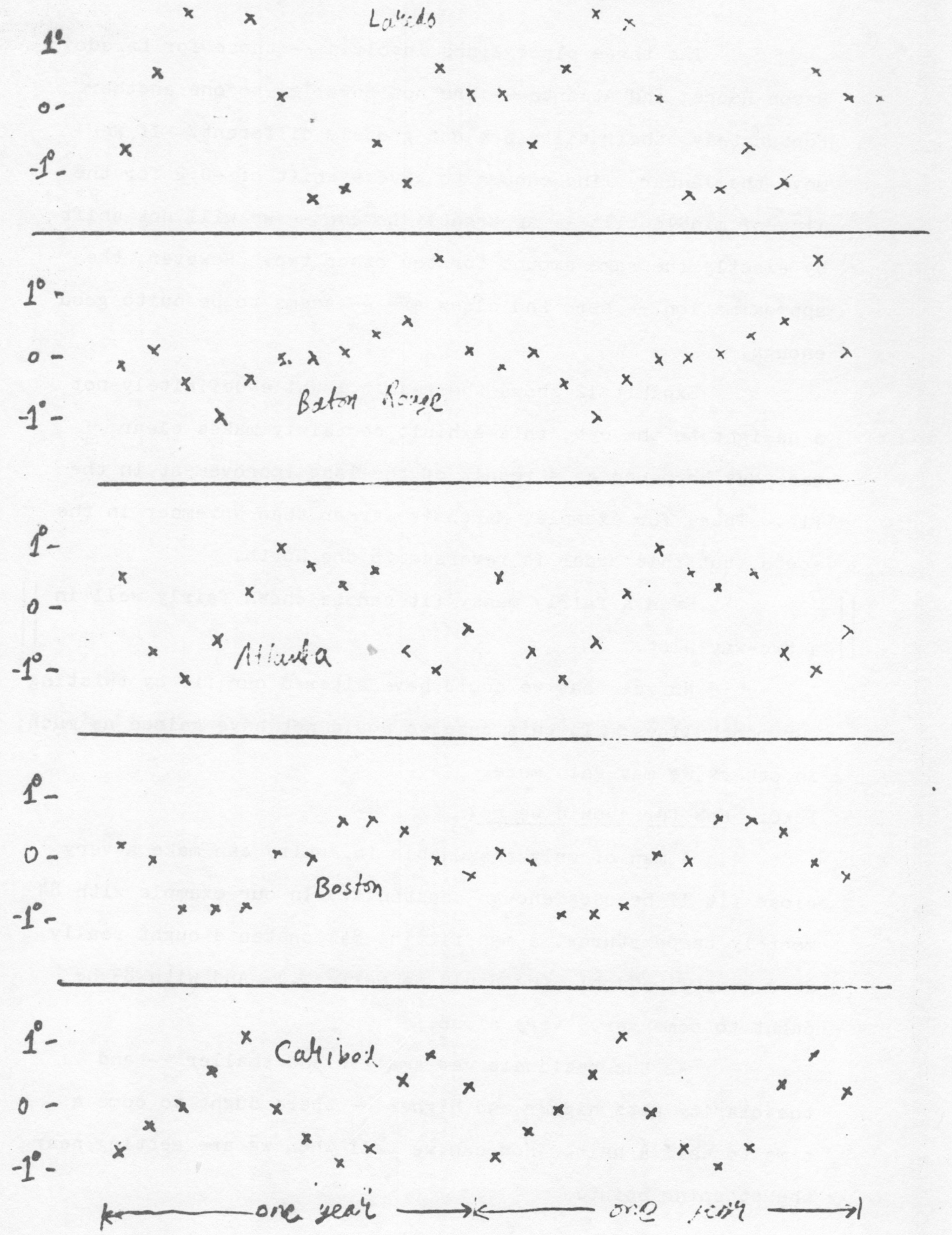
If we want to show our present fit exactly with a two-way plot, it does not seem that we can do it with whole lines -- or even with line segments. If we move the half of the January for "south of Washington" down by 0.2 units, the shift of the intersections will be by somewhat less than 0.2 units. This would be so because the "place" lines are not at right angles to the month lines. (See exhibit 4.)

exhibit 10 of chapter 18x

Uncondensed Residuals
(see exhibit 9)



Residuals Against Jan to Dec (twice)
(see exhibit 9)



The three place lines involved -- those for Laredo, Baton Rouge, and Atlanta -- are not parallel to one another. Fortunately, their tilts are not grossly different. If we move the January line enough to give a shift of -0.2 for the line of middle tilt -- apparently Laredo -- we will not shift by exactly the same amount for the other two. However, the approximation -- here and elsewhere -- seems to be quite good enough.

Exhibit 12 shows the result. While definitely not a delight to the eye, this exhibit certainly makes clear what has happened as a result of the last improvement in the fit. Thus, for example, March is warmer than November in the south, but this order is reversed in the North.

Even a fairly messy fit can be shown fairly well in a two-way plot.

Notice that we could have altered our fit by twisting our month lines. In this case we would not have gained as much; in others we may gain more.

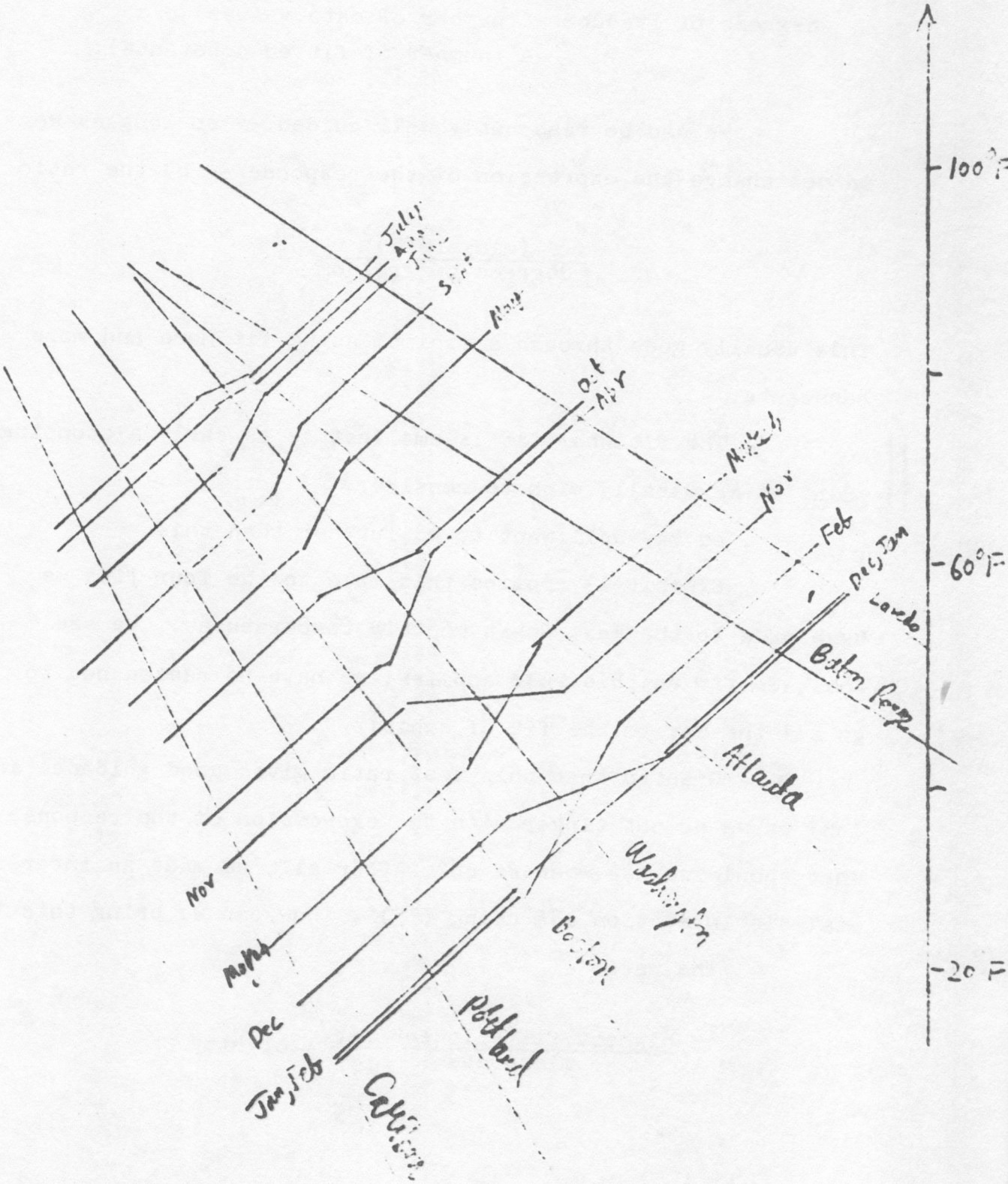
18xC. <u>How far should we go</u>?

A man of only reasonable ingenuity can make a very close fit if he uses enough constants. In our example with 84 monthly temperatures, a man fitting 84 constants ought really do a very fine job! (He could be perfect.) And with 83 he ought to come very, very close.

As the residuals get smaller and smaller -- and the clarity gets higher and higher -- there ought to come a time to call a halt. How can we tell when we are getting near the stopping point?

exhibit 12 of chapter 18x

The Fit of Exhibit 9 as a two-way plot
(approximate, see also exhibit 4)



A handy number to use in such situations is

degrees of freedom = (number of data values)
- (number of fitted constants)

We can be reasonably well guided -- so long as we do not change the expression of the response -- by the ratio

$$\frac{(\text{midspread})^2}{\text{degrees of freedom}}$$

This usually goes through a minimum as we fit more and more constants.

The fit where it is smallest is as early a stopping point as we usually wish to consider.

We may well want to go further than this.

Exhibit 13 applies this rule to the four fits we have made to the East Coast monthly temperatures. We see that, so far as this rule applies, we have no reason not to go all the way to the fit of exhibit 9.

Granted that this test ratio gives good guidance so long as we do not tinker with the expression of the response, what should we use when we do? After all, we made an interesting analysis in section 18A using $(°F)^2$. How can we bring this in?

The ratio

$$\frac{\text{highest fit} - \text{lowest fit}}{\text{midspread}} = \text{clarity}$$

exhibit 13 of chapter 18x

Test Ratios for Four Fits to the
East Coast Monthly Temperatures

Fit No.	1	2	3	4
Fitted	Row effects Column effects	same same (*)	same same (**)	same same (***)
Exhibit	22 of 18	25 of 18	1	9
Constants fitted	12+7-1=18	18+1=19	7+7+10=24	24+12=36
Degrees of freedom remaining	66	65	60	48
Midspread of Residuals	3.7h	1.8h	1.4h	0.9
$\frac{(\text{midspread})^2}{(\text{degrees of freedom})}$	.2	.052	.035	.019

(*) = and a fraction of (row effect)(column effect)

(**) = and separate fractions of seasonal pattern for each place

(***) = as (**) plus a "S of Washington vs. N of Washington" constant for each month.

was introduced to deal with some aspects of this problem. The same philosophy that led to the test ratio just considered leads to

$$(\text{clarity})^2 \cdot \left(\frac{\text{degrees of freedom}}{\text{number of data values}}\right)$$

as an adjusted indication of fit quality.

Exhibit 14 applies this to all five of the fits we have made to the East Coast monthly temperatures. The correction factors are close enough to unity as not to make any notable change in the apparent relative quality of the various fits as judged from

$$(\text{clarity})^2$$

alone.

We can have guidance about how analyses compare -- even when they involve different expressions for the response. Thus we can be guided as to when to stop re-expressing as well as when to stop complicating our fit.

exhibit 14 of chapter 18x

Relative Quality of Five Fits
(East Coast Monthly Temperatures)

	Fits to °F				Fit to $(°F)^2$
	(1)	(2)	(3)	(4)	(5)
Exhibit	22 of 18	25 of 18	1	9	1 of 18
Fitted	row col	row col (*)	row col (**)	row col (***)	row col (power)
df = Degrees of freedom	66	65	60	48	65****
Same/84	.785	.774	.713	.571	.774
Clarity	20.6	42	55	89	31
$(\text{Clarity})^2$	412	1764	3025	7921	961
$(\text{Clarity})^2 \cdot \left(\frac{\text{df}}{\text{count}}\right)$	320	1360	2160	4520	740

(*),(**),(***) are the same as for exhibit 13.

**** We have somewhat overcompensated by subtracting one whole degree of freedom for the choice of the 2nd power in this analysis. (Since we should not have considered all possible powers, an allowance of 2/3 degrees of freedom would have been closer.)

* looking deeper *

Exhibit 15 plots, on a log scale, the figures of merit found in the bottomeline of exhibit 14. Their dependence on degrees of freedom seems reasonable, and the suggestion is that we can expect further improvement from fitting of a sort usable for any two-way table.

We ought not to close our analysis of this example, however, without pointing out that using the specific nature of this data will let us do even better, perhaps coming out in the neighborhood of the "?" on exhibit 15. Our data is monthly data of the kind that repeats -- is periodic. The simplest periodic expressions are not made up of straight lines or quadratics. Rather they are made up of sines and cosines.

For the special case of New England climatic normals, C. Bingham has recently shown how well cosines and sines allow us to fit such data. In view of this, and of exhibit 11, it seems that fitting

$$A + B \cdot \cos 30k^\circ + C \cdot \sin 30k^\circ + D \cdot \cos 60k^\circ + E \cdot \sin 60k^\circ$$

where k = 0 for January, 1 for February, 2 for March, 3 for April, 4 for May, 5 for June, 6 for July, 7 for August, 8 for September, 9 for October, 10 for November, 11 for December --

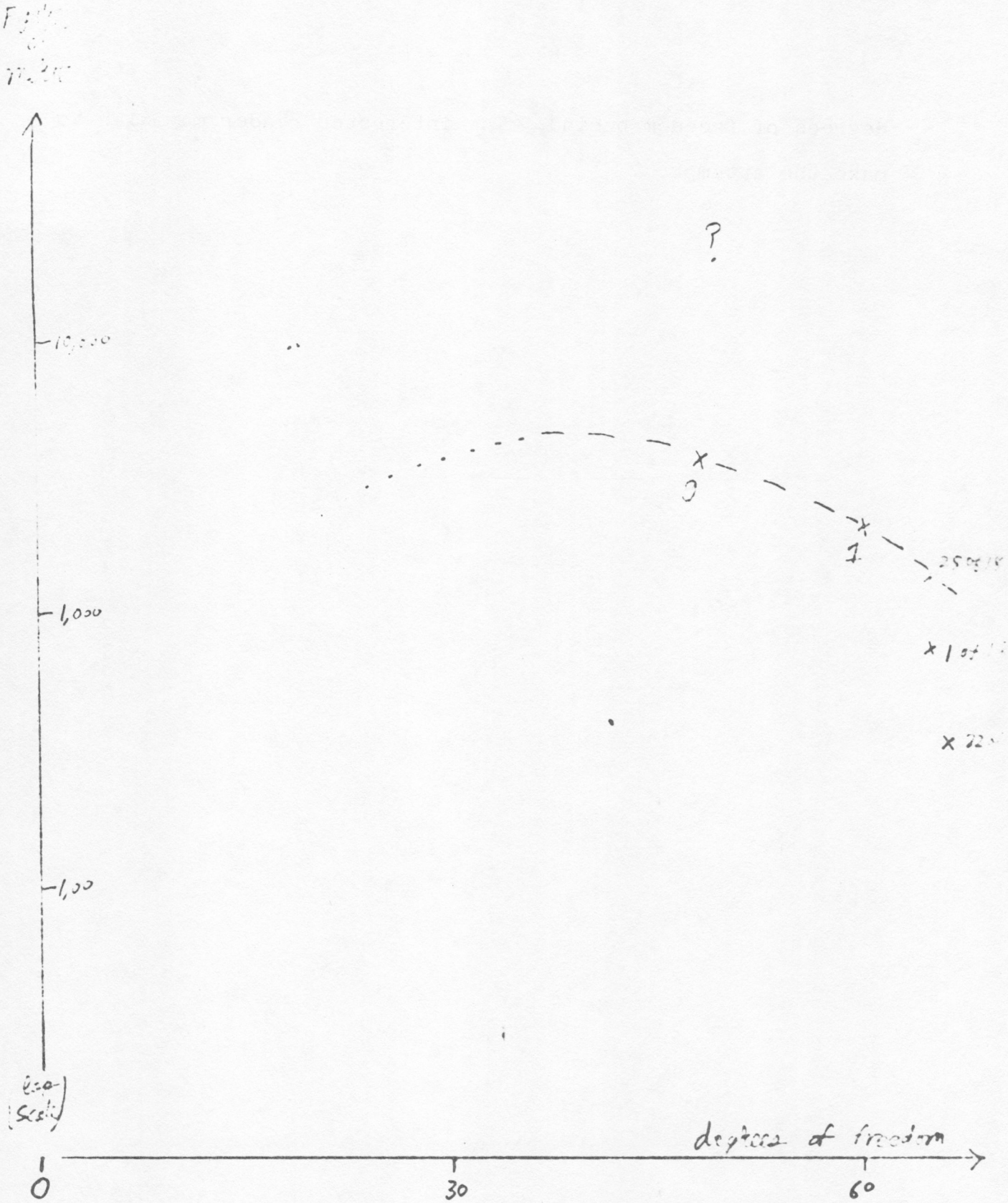
with separate A's, B's, C's, D's and E's for each of the 7 places will do very well indeed. Such a fit leaves

$$84 - 7x5 = 49$$

degrees of freedom behind. The interested reader may wish to make the attempt.

exhibit 15 of chapter 18x

Last line of exhibit 14 plotted against degrees of freedom



******

18xH. <u>How far have we come?</u>

The first step taken in this chapter, in continuing the analysis of chapter 18, was to expand the elementary analysis fit more flexibly, adding either -- but NOT both -- of

(row constant)·(column effect)

or

(column constant)·(row effect)

where, in each instance, the effects were furnished by the elementary analysis and the coefficients are calculated from the elementary-analysis residuals. We added the product of a month constant times the place effect (already found).

Up to this point, we can still make two-way plots of our fitted values using only straight lines.

The next step was to contrast places north of Washington with those south of Washington, separately for each month. Doing this arithmetically was easy, but a two-way plot of the fit required broken lines and was only an approximation though a quite satisfactory one. This is one example of fitting a multiple of a selected carrier to each separate row -- or each separate column. Once a suitable carrier is identified, we can do this for any two-way table.

Alternatively we could have fitted an additional term of the earlier sort, adding the product of a place constant and the month effect (already found).

Once we have a sequence of fits, we need to ask where to stop. (If we fit many constants, we may overfit.)

As long as we keep the same expression for the response, we can judge the relative quality of our fits fairly well from the values of

$$\text{midspread of residuals}$$

or

$$(\text{midspread of residuals})^2$$

small values being desirable. More constants should give better fits, however, even when they shouldn't be used. Useful guidance is offered by

$$\frac{(\text{midspread of residuals})^2}{(\text{degrees of freedom})}$$

where the denominator is equal -- here, as in many simple situations -- to the number of given values minus the number of fitted constants. (We like small values for this ratio!)

When we wish either to deal with more than one way of expressing the response or to put quality on a scale more nearly comparable from one two-way table to another, we can look at the value of

$$\text{clarity} = \frac{\text{highest fit} - \text{lowest fit}}{\text{midspread of residuals}}$$

or at the values of

$$(\text{clarity})^2$$

with large values preferred. When we want to allow for the number of constants fitted, we can use

$$(\text{clarity})^2 \cdot \frac{(\text{degrees of freedom})}{(\#\text{ of row-column combinations}}$$

instead.